Published online 23 October 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D333–D337***
doi:10.1093/nar/gks998

The Online Protein Processing Resource (TOPPR): a database and analysis platform for protein processing events

Niklaas Colaert[1,2], Davy Maddelein[1,2], Francis Impens[1,2], Petra Van Damme[1,2], Kim Plasman[1,2], Kenny Helsens[1,2], Niels Hulstaert[1,2], Joël Vandekerckhove[1,2], Kris Gevaert[1,2,*] and Lennart Martens[1,2,*]

[1]Department of Medical Protein Research, VIB and [2]Department of Biochemistry, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium

Received August 13, 2012; Revised September 27, 2012; Accepted September 28, 2012

ABSTRACT

We here present The Online Protein Processing Resource (TOPPR; http://iomics.ugent.be/toppr/), an online database that contains thousands of published proteolytically processed sites in human and mouse proteins. These cleavage events were identified with COmbinded FRActional DIagonal Chromatography proteomics technologies, and the resulting database is provided with full data provenance. Indeed, TOPPR provides an interactive visual display of the actual fragmentation mass spectrum that led to each identification of a reported processed site, complete with fragment ion annotations and search engine scores. Apart from warehousing and disseminating these data in an intuitive manner, TOPPR also provides an online analysis platform, including methods to analyze protease specificity and substrate-centric analyses. Concretely, TOPPR supports three ways to retrieve data: (i) the retrieval of all substrates for one or more cellular stimuli or assays; (ii) a substrate search by UniProtKB/Swiss-Prot accession number, entry name or description; and (iii) a motif search that retrieves substrates matching a user-defined protease specificity profile. The analysis of the substrates is supported through the presence of a variety of annotations, including predicted secondary structure, known domains and experimentally obtained 3D structure where available. Across substrates, substrate orthologs and conserved sequence stretches can also be shown, with iceLogo visualization provided for the latter.

More than two percent of all human and mouse genes encode proteases. These enzymes control many biological processes and are of crucial importance for relatively simple processes such as food digestion, as well as for highly regulated proteolytic cascades such as controlled cell death or blood coagulation. In addition, misregulated protease activities add to the severity of several pathologies, including cancer, cardiovascular and inflammatory diseases. It is commonly recognized that a more detailed understanding of protease-controlled or protease-affected processes can be achieved by extending our overall knowledge on proteases, their (preferred) substrates and specificities (1).

The N-terminal COFRADIC (COmbinded FRActional DIagonal Chromatography) technique developed in our laboratory enables isolation and identification of protein N-terminal peptides using peptide chromatography and mass spectrometry (MS) (2). Given that protein processing induces new N- and C-terminal protein ends, the resulting neo-N-terminal peptides are also isolated and identified, and represent proxies for the actual cleavage position in the protease substrate (3). Recently, a similar C-terminal COFRADIC technique was developed to select and identify (neo-) C-terminal peptides, thus also identifying processing events (4). These COFRADIC techniques made the identifications of large numbers of processing events possible, and these techniques are applicable for both individual proteases and in cellular setups in which several proteases are active (3,5–12). It should be noted that besides the COFRADIC technologies, other mass

*To whom correspondence should be addressed. Tel: +32 9 264 9274; Fax: +32 9 264 9496; Email: kris.gevaert@vib-ugent.be
Correspondence may also be addressed to Lennart Martens. Tel: +32 9 264 9358; Fax: +32 9 264 9484; Email: lennart.martens@vib-ugent.be

The authors wish to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

spectrometry-based technologies were recently introduced to identify protein processing events [recently reviewed in (13)].

A number of databases are currently available that disseminate protein processing events. The MEROPS database is specialized in proteases and their classification, and stores substrates linked to proteases (14). However, this database does not easily allow a user to perform specific meta-analyses such as a direct comparison among substrates. PMAP/CutDB on the other hand is a community-driven (Wikipedia style) database, implying that any scientist can add new substrates or substrate predictions (15,16). The intrinsic disadvantage is that the quality of the reported substrates cannot be guaranteed, especially because the original data leading to the discovery of a substrate can only be accessed by using the provided links to the original article. CASBAH, developed in 2007 as part of a review article on proteolytic processes in dying cells, stores processing sites of caspases only (17). It is also important to note that a processing event in a substrate stored in CutDB or CASBAH is not necessarily linked to the actual cleavage position in the substrate. Recently, the TopFIND 2.0 database has been released as well, offering a protein-centric knowledgebase on protein termini, including modifications and processing events (18). TopFIND comes equipped with data mining and analysis tools, but is ultimately based on curated (imported) data from experimental data sets and existing databases, losing any direct connection to the underlying experimental data in the process. As such, it resembles the UniProtKB/Swiss-Prot model, albeit with more integrated analysis tools. Finally, the ApoptoProteomics database was also recently launched, but as can be derived from its name, this database centralizes on protein processing found in apoptotic cells from different origins (19).

In conclusion, no single database exists today that provides complete data provenance, an essential feature in guaranteeing data quality, with only TopFIND 2.0 and ApoptoProteomics providing support for further (meta-) analyses on both proteases and their substrates.

We here present The Online Protein Processing Resource (TOPPR) that stores published processing events identified in our laboratory by N- and C-terminal COFRADIC technologies. TOPPR makes our data available through an easy and intuitive analysis platform. Furthermore, the application provides a user interface that is specifically tailored to verify the actual MS/MS data that led to the identification of the reported processed sites. In fact, the Mascot identification score (20), corresponding threshold score and confidence level are easily checked for every peptide reporting a processing event. Additionally, the b- and y-ion annotated MS/MS spectrum can be viewed and downloaded using the PRIDE spectrum viewer (21). These annotations are derived from the underlying ms_lims processing pipeline (22) that is in turn built on the MascotDatfile library for reading and interpreting Mascot search results (23). Full data provenance is thus guaranteed, making it simple for the user to check data quality. Note that this implies that TOPPR is exclusively dedicated to displaying experimentally observed cleavages sites, and that the system does not include any predicted cleavage sites. The focus on empirical mass spectrometry-based proteomics data also means that TOPPR will under-represent any cleavages sites that are difficult to detect with this technology, notably in the case of heavily modified peptides. TOPPR also supports user-level security, allowing data to be kept private to one or more authenticated users before publication. All information in TOPPR is stored in a MySQL database (see Supplementary Figure S1 for the relational schema), and query results are generated by JavaServer Pages and Java Servlet technologies running on an Apache Tomcat server infrastructure. TOPPR is released under the permissive Apache2 open source license, and all source code can be downloaded from http://code.google.com/p/toppr/.

At the time of writing, TOPPR contains 2234 substrates, for 18 studied treatments or peptidases, resulting in 27 147 cleavages. To navigate these data, TOPPR provides three different search methods. The first method, the parameter search, is used to find all substrates for one treatment or a combination of treatments, where a treatment corresponds to either a cellular stimulus in an *in vivo*/*in cellulo* assay or, alternatively, a protease used in an *in vitro* assay (i.e. a protease added to a cell lysate). The user selects one or more treatments from a list of published treatments, and can perform a range of set operations on these to create specific queries. Furthermore, this query interface allows even more fine-grained retrieval options to be specified, including combinations of treatments that result in the same site being cleaved (processing 'hot spots'). Second, a UniProtKB/Swiss-Prot search is provided by which users can use a UniProtKB/Swiss-Prot accession number, a UniProtKB/Swiss-Prot entry name or a fraction of the corresponding protein description as the search string. This method will reveal all stored processed sites linked to the specified substrate. Third, a motif search enables users to search for substrates containing processed sites that match the user-defined protease specificity profile. This specificity profile is defined in two parts: a pre-site (non-primed sites) and a post-site motif (primed sites) (24), with each motif defined using a simplified regular expression syntax.

TOPPR also supports two types of analysis: analysis of the processed sites and corresponding protease specificity and, in addition, detailed analysis of individual substrates. The analysis of processed sites to infer protease specificity can be carried out by using integrated tools like iceLogo [probability-based visualization of significantly enriched/depleted residues in aligned sequences (25)], Weblogo [sequence logos (26)], PoPS [prediction of protease specificity (27)] and JalView [multiple sequence viewer and analysis tool (28)]. The list of processed sites used as input for these tools is extracted from TOPPR through the data retrieval options listed earlier, or as a manually selected subset of sites. The detailed analysis of individual substrates, on the other hand, is provided in TOPPR through a variety of integrated substrate metadata whenever available. First of all, processing events by different proteases are readily visualized using the substrate sequence view of TOPPR (see Figure 1). Additionally, Smart (29) and Pfam (30) annotations can be shown

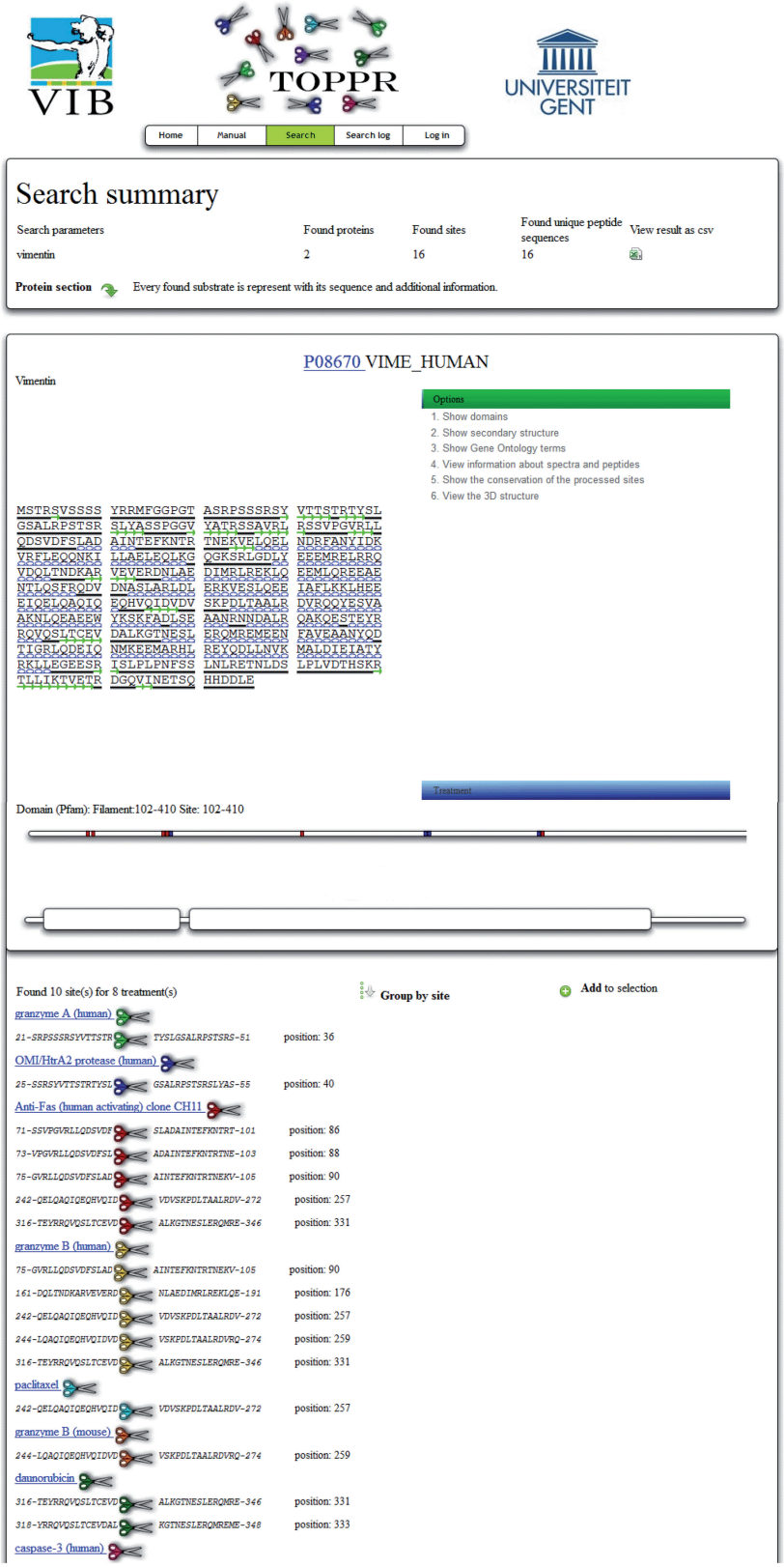

Figure 1. The substrate sequence view in TOPPR. This display provides an overview of the protein sequence and links to dynamic annotations (here secondary structure and domain annotation have been selected). A protein bar representation, below the sequence, represents the full length of the protein with processing events indicated; immediately underneath, the domain visualization is shown. Below this protein-centric sequence view, details are shown on the individual peptides that were found to represent the annotated cleavage sites. Each peptide sequence in turn links to a peptide-centric view, where motif analyses, mass spectrometry data and all matching proteins can be found. Note that the peptide is annotated with its start and end coordinates on the protein.

alongside reported processing events at the substrate level, thus indicating possible processing-derived interference with substrate function. Conservation of processing events among different species can be studied via a built-in function that globally aligns the surrounding sequences of processed sites in all known substrate homologs or orthologs (found in the HomoloGene database). Where available, TOPPR provides 3D structures (visualized with JMol) of the substrates. Processing events are indicated in these structures using a ball and stick configuration, whereas the rest of the polypeptide chain is in the cartoon configuration. This visualization makes it straightforward to assess the processing event in the context of the substrate's 3D structure. Finally, the UniProtKB/Swiss-Prot annotated secondary structure elements, or, if unavailable, a secondary structure prediction (31) can be shown in the sequence view, facilitating substrate examination in the absence of 3D structures.

The TOPPR database is continuously updated with novel findings, keeping track of all published protease cleavage sites identified by COFRADIC (or related) technologies in our laboratory. Over time, more treatments and substrates will therefore be included, leading to an increasingly comprehensive database of cleavage sites for the most abundant eukaryotic intracellular proteases (e.g. human and mouse caspases, granzymes, calpains and cathepsins). Furthermore, through the ability to transmit the data associated with an entire project from one ms_lims system to another via the Internet, TOPPR can easily receive incoming data from third parties. Users of ms_lims need only contact the authors for a username and password to connect to the TOPPR-linked ms_lims installation at the authors' laboratories, at which point the project transmission application of ms_lims allows the data to be transmitted with a single click, ensuring its downstream uptake in TOPPR as well. Note that the reliance on ms_lims as the underlying data processing and management platform implicitly ensures consistency and comparable quality across all assembled data. Apart from its role as a data storage system, TOPPR also serves as a powerful exploration platform to verify data quality, assess protease specificity, perform processing site motif analyses and carry out detailed substrate analyses. An online user manual provides detailed information on the available search methods and types of analysis. TOPPR thus provides a powerful platform for discovery and analysis to both protease researchers and scientists studying a single substrate.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figure 1.

FUNDING

Ghent University Multidisciplinary Research Partnership 'Bioinformatics: from nucleotides to networks'; Fund for Scientific Research (FWO)—Flanders (Belgium) (postdoctoral research fellowship to F.I., P.V.D. and K.H.); Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) (PhD to K.P.); ProteomeXchange project, funded by the European Union 7th Framework Program under grant agreement number [260558 to N.H.]; PRIME-XS project, funded by the European Union 7th Framework Program under grant agreement number [262067 to K.G. and L.M.]. Funding for open access charge: VIB, Ghent, Belgium.

Conflict of interest statement. None declared.

REFERENCES

1. Puente,X.S., Sánchez,L.M., Overall,C.M. and López-Otín,C. (2003) Human and mouse proteases: a comparative genomic approach. *Nat. Rev. Genet.*, **4**, 544–548.
2. Gevaert,K., Goethals,M., Martens,L., Van Damme,J., Staes,A., Thomas,G. and Vandekerckhove,J. (2003) Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat. Biotechnol.*, **21**, 566–569.
3. Van Damme,P., Martens,L., Van Damme,J., Hugelier,K., Staes,A., Vandekerckhove,J. and Gevaert,K. (2005) Caspase-specific and nonspecific in vivo protein processing during Fas-induced apoptosis. *Nat. Methods*, **2**, 771–777.
4. Van Damme,P., Staes,A., Bronsoms,S., Helsens,K., Colaert,N., Timmerman,E., Aviles,F.X., Vandekerckhove,J. and Gevaert,K. (2010) Complementary positional proteomics for screening substrates of endo- and exoproteases. *Nat. Methods*, **7**, 512–515.
5. Vande Walle,L., Van Damme,P., Lamkanfi,M., Saelens,X., Vandekerckhove,J., Gevaert,K. and Vandenabeele,P. (2007) Proteome-wide identification of HtrA2/Omi substrates. *J. Proteome Res.*, **6**, 1006–1015.
6. Impens,F., Van Damme,P., Demol,H., Van Damme,J., Vandekerckhove,J. and Gevaert,K. (2008) Mechanistic insight into taxol-induced cell death. *Oncogene*, **27**, 4580–4591.
7. Lamkanfi,M., Kanneganti,T.-D., Van Damme,P., Vanden Berghe,T., Vanoverberghe,I., Vandekerckhove,J., Vandenabeele,P., Gevaert,K. and Núñez,G. (2008) Targeted peptidecentric proteomics reveals caspase-7 as a substrate of the caspase-1 inflammasomes. *Mol. Cell. Proteomics*, **7**, 2350–2363.
8. Van Damme,P., Maurer-Stroh,S., Plasman,K., Van Durme,J., Colaert,N., Timmerman,E., De Bock,P.-J., Goethals,M., Rousseau,F., Schymkowitz,J. *et al.* (2008) Analysis of protein processing by N-terminal proteomics reveals novel species-specific substrate determinants of granzyme B orthologs. *Mol. Cell. Proteomics*, **8**, 258–272.
9. Demon,D., Van Damme,P., Vanden Berghe,T., Deceuninck,A., Van Durme,J., Verspurten,J., Helsens,K., Impens,F., Wejda,M., Schymkowitz,J. *et al.* (2009) Proteome-wide substrate analysis indicates substrate exclusion as a mechanism to generate caspase-7 versus caspase-3 specificity. *Mol. Cell. Proteomics*, **8**, 2700–2714.
10. Impens,F., Colaert,N., Helsens,K., Ghesquiere,B., Timmerman,E., De Bock,P.-J., Chain,B.M., Vandekerckhove,J. and Gevaert,K. (2010) A quantitative proteomics design for systematic identification of protease cleavage events. *Mol. Cell. Proteomics*, **9**, 2327–2333.
11. Van Damme,P., Maurer-Stroh,S., Hao,H., Colaert,N., Timmerman,E., Eisenhaber,F., Vandekerckhove,J. and Gevaert,K. (2010) The substrate specificity profile of human granzyme A. *Biol. Chem.*, **391**, 983–997.
12. Kaiserman,D., Buckle,A.M., Van Damme,P., Irving,J.A., Law,R.H.P., Matthews,A.Y., Bashtannyk-Puhalovich,T., Langendorf,C., Thompson,P., Vandekerckhove,J. *et al.* (2009) Structure of granzyme C reveals an unusual mechanism of protease autoinhibition. *Proc. Natl Acad. Sci. USA*, **106**, 5587–5592.
13. Impens,F., Vandekerckhove,J. and Gevaert,K. (2010) Who gets cut during cell death? *Curr. Opin. Cell Biol.*, **22**, 859–864.

14. Rawlings,N.D., Barrett,A.J. and Bateman,A. (2010) MEROPS: the peptidase database. *Nucleic Acids Res.*, **38**, D227–D233.
15. Igarashi,Y., Eroshkin,A., Gramatikova,S., Gramatikoff,K., Zhang,Y., Smith,J.W., Osterman,A.L. and Godzik,A. (2007) CutDB: a proteolytic event database. *Nucleic Acids Res.*, **35**, D546–D549.
16. Igarashi,Y., Heureux,E., Doctor,K.S., Talwar,P., Gramatikova,S., Gramatikoff,K., Zhang,Y., Blinov,M., Ibragimova,S.S., Boyd,S. *et al.* (2009) PMAP: databases for analyzing proteolytic events and pathways. *Nucleic Acids Res.*, **37**, D611–D618.
17. Lüthi,A.U. and Martin,S.J. (2007) The CASBAH: a searchable database of caspase substrates. *Cell Death Differ.*, **14**, 641–650.
18. Lange,P.F., Huesgen,P.F. and Overall,C.M. (2012) TopFIND 2.0—linking protein termini with proteolytic processing and modifications altering protein function. *Nucleic Acids Res.*, **40**, D351–D361.
19. Arntzen,M.Ø. and Thiede,B. (2012) ApoptoProteomics, an integrated database for analysis of proteomics data obtained from apoptotic cells. *Mol. Cell. Proteomics*, **11**, M111.010447.
20. Perkins,D.N., Pappin,D.J.C., Creasy,D.M. and Cottrell,J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.
21. Vizcaíno,J.A., Côté,R., Reisinger,F., Foster,J.M., Mueller,M., Rameseder,J., Hermjakob,H. and Martens,L. (2009) A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics*, **9**, 4276–4283.
22. Helsens,K., Colaert,N., Barsnes,H., Muth,T., Flikka,K., Staes,A., Timmerman,E., Wortelkamp,S., Sickmann,A., Vandekerckhove,J. *et al.* (2010) ms_lims, a simple yet powerful open source laboratory information management system for MS-driven proteomics. *Proteomics*, **10**, 1261–1264.
23. Helsens,K., Martens,L., Vandekerckhove,J. and Gevaert,K. (2007) MascotDatfile: an open-source library to fully parse and analyse MASCOT MS/MS search results. *Proteomics*, **7**, 364–366.
24. Schechter,I. and Berger,A. (1967) On the size of the active site in proteases. I. Papain. *Biochem. Biophys. Res. Commun.*, **27**, 157–162.
25. Colaert,N., Helsens,K., Martens,L., Vandekerckhove,J. and Gevaert,K. (2009) Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods*, **6**, 786–787.
26. Crooks,G.E., Hon,G., Chandonia,J.-M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
27. Boyd,S., Pike,R., Rudy,G., Whisstock,J. and Garcia De La Banda,M. (2005) PoPS: a computational tool for modeling and predicting protease specificity. *J. Bioinform. Comput. Biol.*, **3**, 551–585.
28. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
29. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
30. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
31. Chandonia,J.-M. (2007) StrBioLib: a Java library for development of custom computational structural biology applications. *Bioinformatics*, **23**, 2018–2020.